Problems of Fuzzy c-Means and similar algorithms with high dimensional data sets

Roland Winkler (roland.winkler@dlr.de), Frank Klawonn, Rudolf Kruse

July 20, 2010

Deutsches Zentrum
für Luft- und Raumfahrt e.V.
in der Helmholtz-Gemeinschaft

Outline

Current Section

Prototype based algorithms

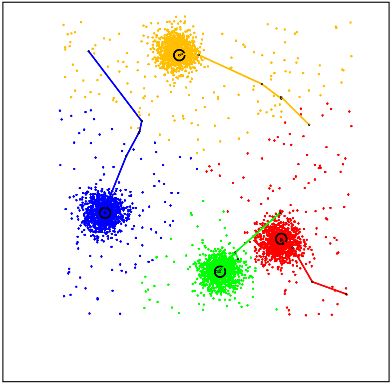
- Objective function that needs to be minimized:

$$J = \sum_{i=1}^{c} \sum_{j=1}^{n} f(u_{ij}) d_{ij}^2$$

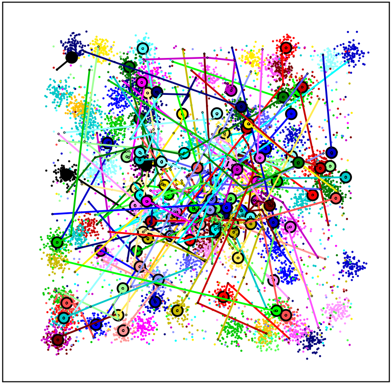
- c: number of prototypes, n: number of data objects, u_{ij} membership value, $f : \mathbb{R} \to \mathbb{R}$: fuzzifier function, $d_{ij} = \|y_i - x_j\|$, y_i: prototype, x_j: data object, $1 = \sum_{i=1}^{c} u_{ij}$
- artificial generated data set 1: 2 dimensions, 4 clusters, 1000 data objects per cluster, 10% noise
- artificial generated data set 2: 50 dimensions, 100 clusters, 100 data objects per cluster, 10% noise

HCM

- $f(u) = u$



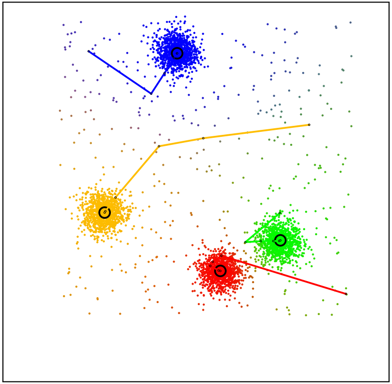
Data set 1

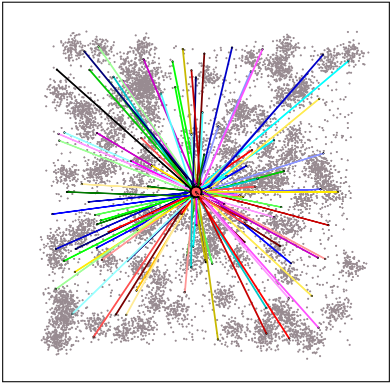


Data set 2, found clusters: 40

FCM

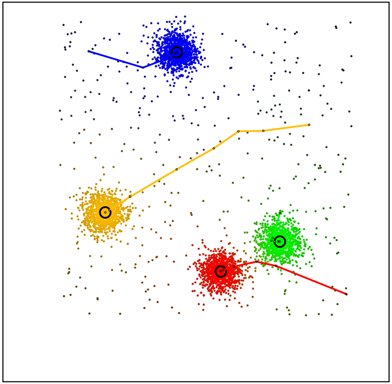
- $f(u) = u^{\omega}$, $\omega = 2$

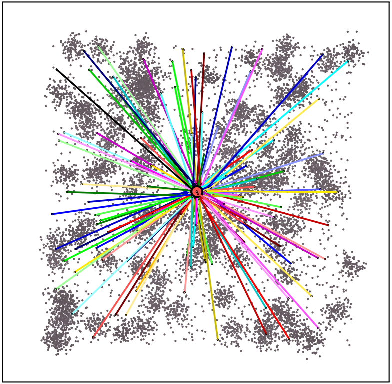

Data set 1


Data set 2, found clusters: 0

NFCM

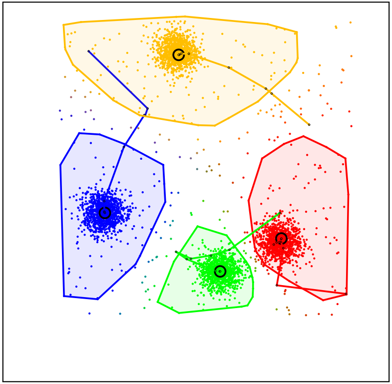
- $f(u) = u^{\omega}$, $\omega = 2$, noise cluster at $d_{\text{noise}} = 0.3$

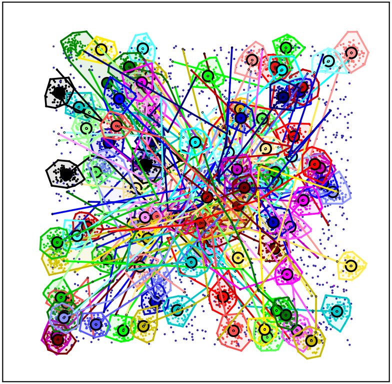

Data set 1


Data set 2, found clusters: 0

PFCM

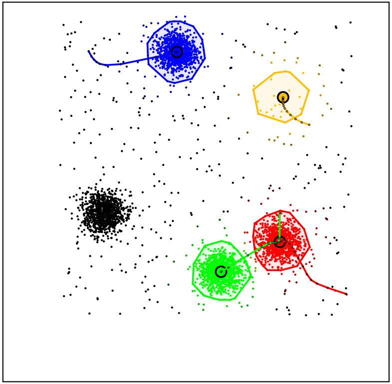
- $f(u) = \left(\frac{1-\beta}{1+\beta}u^2 + \frac{2\beta}{1+\beta}u\right)$, $\beta = 0.5$
- in a 2 cluster environment: $u_{1j} = 1 \Leftrightarrow \frac{d_{1j}}{d_{2j}} < \beta$


Data set 1

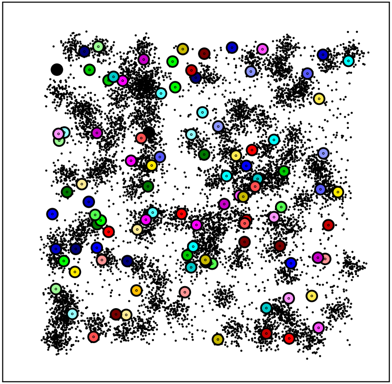

Data set 2, found clusters: 90

PNFCM

- $f(u) = \left(\frac{1-\beta}{1+\beta}u^2 + \frac{2\beta}{1+\beta}u\right)$, $\beta = 0.5$, $d_{\text{noise}} = 0.3$



Data set 1

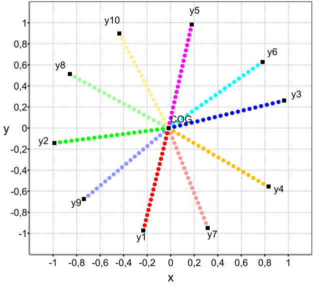


Data set 2, found clusters: 0

Current Section

Test setup

- Generate a data set with $n = c$ data objects representing c clusters of infinite density
- Place the data objects on a D-dimensional hypersphere surface of radios 1 such that the pairwise distances are maximised
- Place prototypes in the centre of gravity (cog) and gradually move them to the data objects
- Observe the objective function

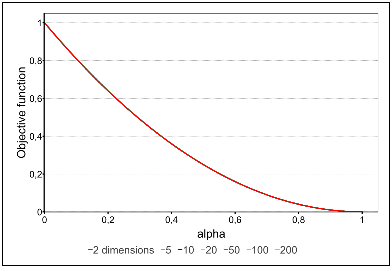


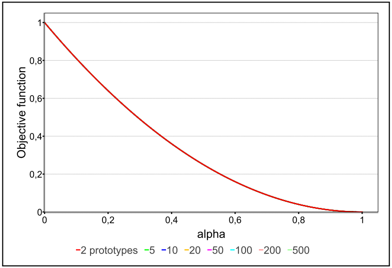
Tests

- Applying the algorithms
  - $y_i = (1 - \alpha) \cdot \text{cog} + \alpha x_i$ for $\alpha \in [0, 1]$
  - Normalise the objective function by the value of the objective function at $\alpha = 0$
  - Let α gradually increase from 0 to 1 and monitor the normalized objective function value
- 2 Test setups
  - Setup 1: 100 prototypes, between 2 and 200 dimensions
  - Setup 2: 50 dimensions, between 2 and 500 prototypes

Current Section

Objective function of HCM



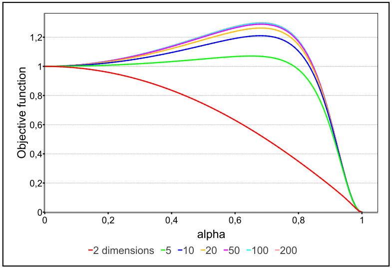


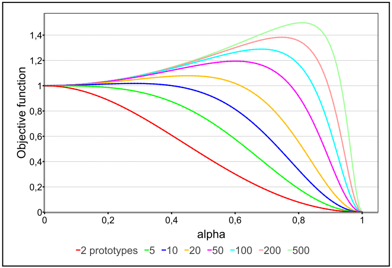
100 prototypes
various dimensions

50 Dimensions
various prototypes

Objective function of FCM

- $f(u) = u^{\omega}, \omega = 2$



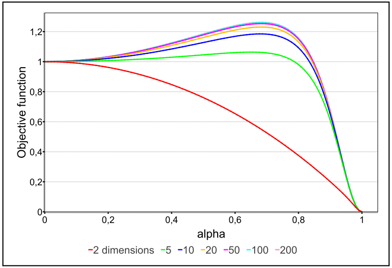


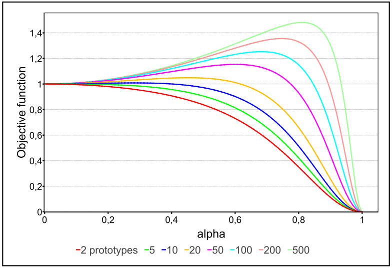
100 prototypes
various dimensions

50 Dimensions
various prototypes

Objective function of NFCM

- $f(u) = u^{\omega}$, $\omega = 2$, $d_{\text{noise}} = \frac{1}{2}$



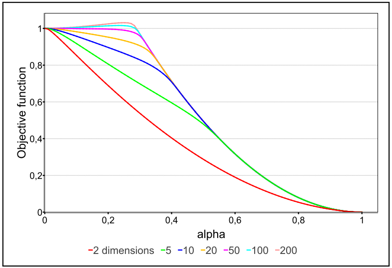


100 prototypes
various dimensions

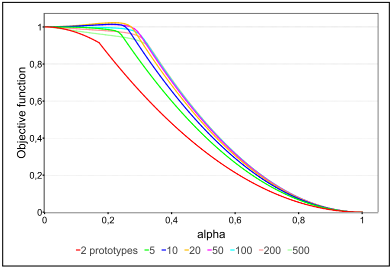
50 Dimensions
various prototypes

Objective function of PFCM

- $f(u) = \left(\frac{1-\beta}{1+\beta}u^2 + \frac{2\beta}{1+\beta}u\right)$, $\beta = \frac{1}{2}$



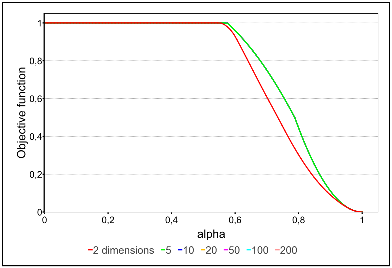
100 prototypes
various dimensions



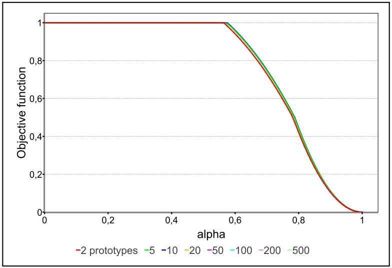
50 Dimensions
various prototypes

Objective function of PNFCM

- $f(u) = \left(\frac{1-\beta}{1+\beta}u^2 + \frac{2\beta}{1+\beta}u\right)$, $\beta = \frac{1}{2}$, $d_{\text{noise}} = \frac{1}{2}$



100 prototypes
various dimensions



50 Dimensions
various prototypes

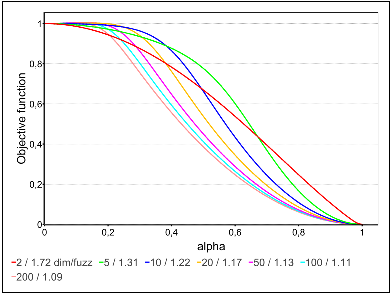
Deutsches Zentrum für Luft- und Raumfahrt e.V.
in der Helmholtz-Gemeinschaft

Tweak the fuzzy based algorithms to work at high dimensions

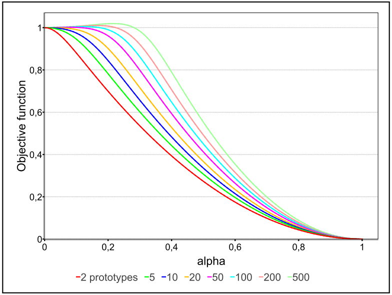
- For FCM, a dimension dependent fuzzifier can be used to counter the dimension effect: $\omega = 1 + \frac{2}{D}$
- With a tweaked fuzzifier, NFCM behaves like PNFCM and needs a dimension dependent noise distance: $d_{\text{noise}} = 0.5 \log_2(D)$
- PNFCM requires also requires an dimension dependent noise distance: $d_{\text{noise}} = 0.5 \log_2(D)$

FCM with adjusted parameters

- $f(u) = u^{\omega}, \omega = 1 + \frac{2}{D}$

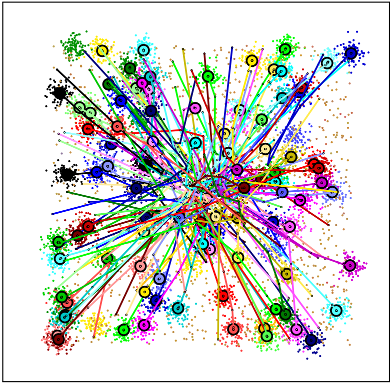


100 prototypes
various dimensions



50 Dimensions
various prototypes

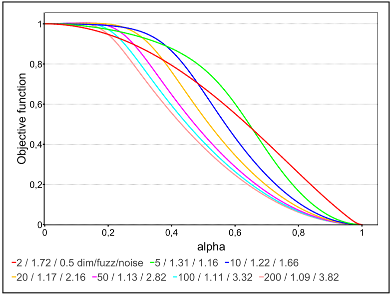
FCM with adjusted parameters



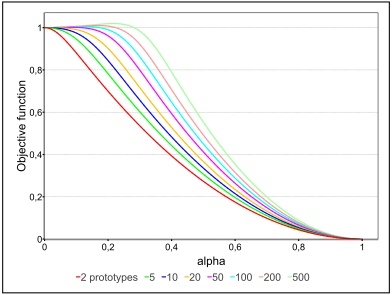
50 Dimensions, 100 clusters, found clusters: 92

NFCM with adjusted parameters

- $f(u) = u^{\omega}$, $\omega = 1 + \frac{2}{D}$, $d_{\text{noise}} = 0.5 \log_2(D)$

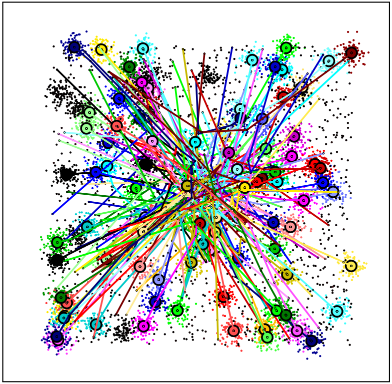


100 prototypes
various dimensions



50 Dimensions
various prototypes

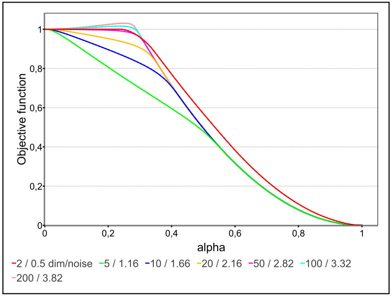
NFCM with adjusted parameters

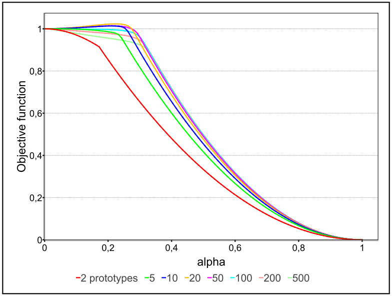


50 Dimensions, 100 clusters, found clusters: 91, correct clustered noise: 1000, incorrect clustered as noise: 900

PNFCM with adjusted parameters

- $f(u) = \left(\frac{1-\beta}{1+\beta}u^2 + \frac{2\beta}{1+\beta}u\right)$, $\beta = 0.5$, $d_{noise} = 0.5\log_2(D)$

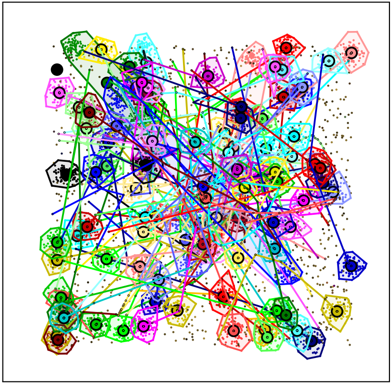




100 prototypes
various dimensions

50 Dimensions
various prototypes

PNFCM with adjusted parameters



50 Dimensions, 100 clusters, found clusters: 90, correct clustered noise: 991, incorrect clustered as noise: 0

Current Section

Scoring the algorithms

- 50 Dimensions, 100 cluster with 100 data objects each, 1000 noise data objects

Algorithm	Found clusters	Correctly clustered noise	Incorrect clustered as noise
HCM	40	0	0
FCM	0	0	0
NFCM	0	1000	10000
PFCM	90	0	0
PNFCM	0	1000	10000
Adjusted Parameter			
FCM AP	92	0	0
NFCM AP	91	1000	900
PNFCM AP	90	994	0

Effects of high dimensions on the feature space

- Let S be a hypersphere of D dimensions with radius R.
- S has a volume of $V = C_D \cdot R^D$ with $C_D = \frac{\pi^{\frac{D}{2}}}{\Gamma\left(\frac{D}{2}+1\right)}$ only depending on D
- Let S^* be a hypersphere of D dimensions with radius R^* and $V^* = C_D \cdot {R^*}^D = \frac{1}{2}V$
- Then $R^* = \left(\frac{1}{2}\right)^{\frac{1}{D}} R$
- Example:

$D = 2$:	$R^* = 0.70711R$	$D = 10$:	$R^* = 0.93303R$
$D = 100$:	$R^* = 0.99309R$	$D = 1000$:	$R^* = 0.99931R$

Current Section

Literature

Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft.
When is nearest neighbor meaningful?
In *Database Theory - ICDT'99*, volume 1540 of *Lecture Notes in Computer Science*, pages 217–235. Springer Berlin / Heidelberg, 1999.

F. Höppner, F. Klawonn, R. Kruse, and T. Runkler.
Fuzzy Cluster Analysis.
John Wiley & Sons, Chichester, England, 1999.

Frank Klawonn and Frank Höppner.
What is fuzzy about fuzzy clustering? understanding and improving the concept of the fuzzifier.
In *Cryptographic Hardware and Embedded Systems - CHES 2003*, volume 2779 of *Lecture Notes in Computer Science*, pages 254–264. Springer Berlin / Heidelberg, 2003.

Thank you for your attention

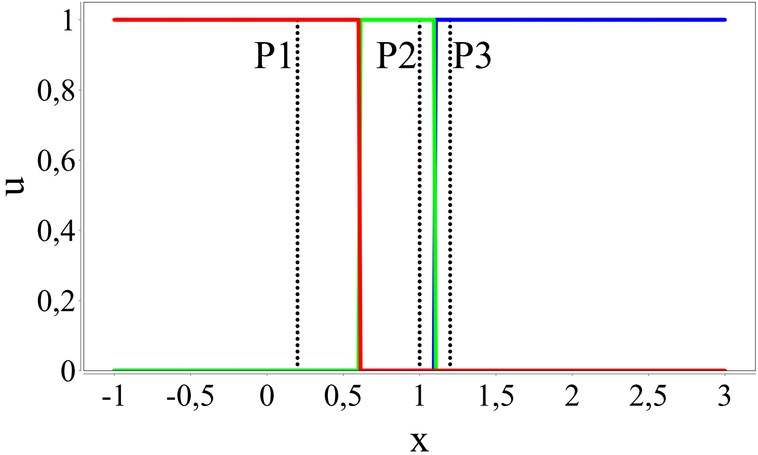
Current Section

Hard c-means (HCM)

- Clustering Problem: data objects $X = \{x_1, \dots, x_n\}$, prototypes $Y = \{y_1, \dots, y_c\}$
- Objective Function: $\boldsymbol{J}_{\text{HCM}} = \min_S \sum_{j=1}^{n} \sum_{j \in S_i} d_{ij}^2$
  with $d_{ij} = \|y_i - x_j\|$, $S = \{S_1, \dots, S_c\}$, $S_i \subset \mathbb{N}$ and $S_i \cap S_k = \emptyset$ for $i \neq k$.
- updated with:
  $$S_i = \left\{ j \ : \ d_{ij} \leq d_{ik} \forall k \right\} \quad \text{and} \quad y_i = \frac{1}{|S_i|} \sum_{j \in S_i} x_j$$

Deutsches Zentrum für Luft- und Raumfahrt e.V. in der Helmholtz-Gemeinschaft

HCM membership values example



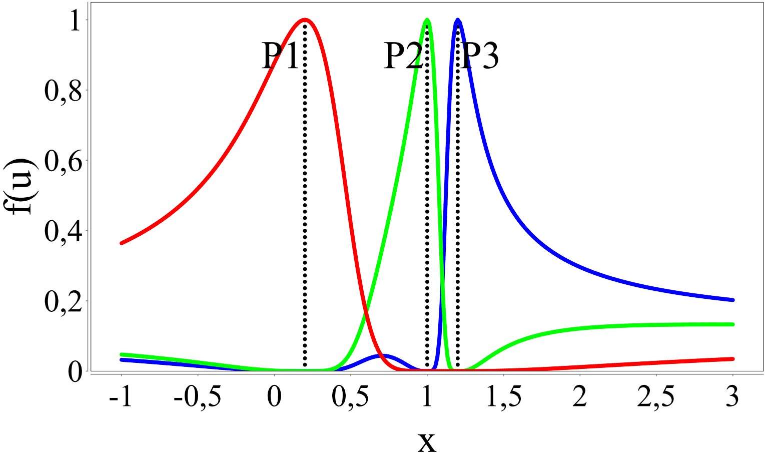
Fuzzy c-means (FCM)

- Clustering Problem: data objects $X = \{x_1, \ldots, x_n\}$, prototypes $Y = \{y_1, \ldots, y_c\}$
- Objective Function: $\boldsymbol{J}_{\text{FCM}} = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^{\omega} d_{ij}^2$

  with $d_{ij} = \|y_i - x_j\|$, $1 = \sum_{i=1}^{c} u_{ij}$, $u_{ij} > 1$ and $\omega > 1$.
- updated with:

$$u_{ij} = \frac{d_{ij}^{\frac{2}{1-\omega}}}{\sum_{k=1}^{c} d_{kj}^{\frac{2}{1-\omega}}} \quad \text{and} \quad y_i = \frac{\sum_{j=1}^{n} u_{ij}^{\omega} \cdot x_j}{\sum_{j=1}^{n} u_{ij}^{\omega}}.$$

FCM fuzzified membership values example

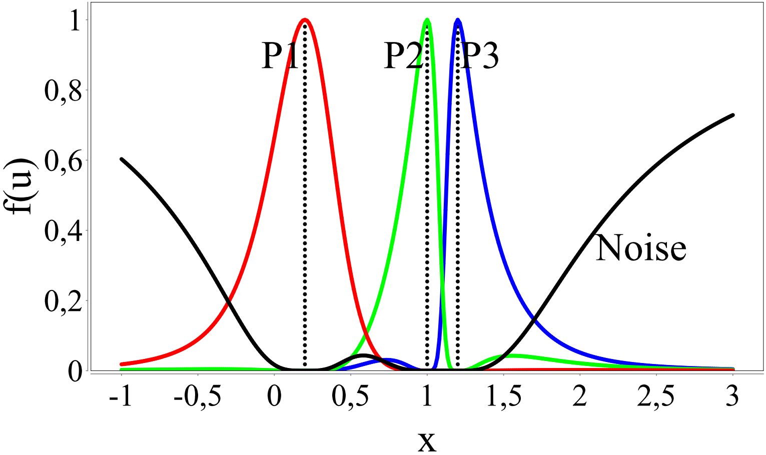


Noise Fuzzy c-means (NFCM)

- Objective Function: $\boldsymbol{J}_{\text{NFCM}} = \sum_{i=0}^{c} \sum_{j=1}^{n} u_{ij}^{\omega} d_{ij}^{2}$
  with $d_{0j} = d_{\text{noise}} \geq 0$, $d_{ij} = \|y_i - x_j\|$ for $i = 1 \ldots c$,
  $1 = \sum_{i=1}^{c} u_{ij}$, $u_{ij} > 1$ and $\omega > 1$.
- updated with:

$$u_{ij} = \frac{d_{ij}^{\frac{2}{1-\omega}}}{\sum_{k=0}^{c} d_{kj}^{\frac{2}{1-\omega}}} \quad \text{and} \quad y_i \overset{i=1\ldots c}{=} \frac{\sum_{j=1}^{n} u_{ij}^{\omega} \cdot x_j}{\sum_{j=1}^{n} u_{ij}^{\omega}}$$

NFCM fuzzified membership values example



Fuzzy c-means with polynomial fuzzifier (PFCM)

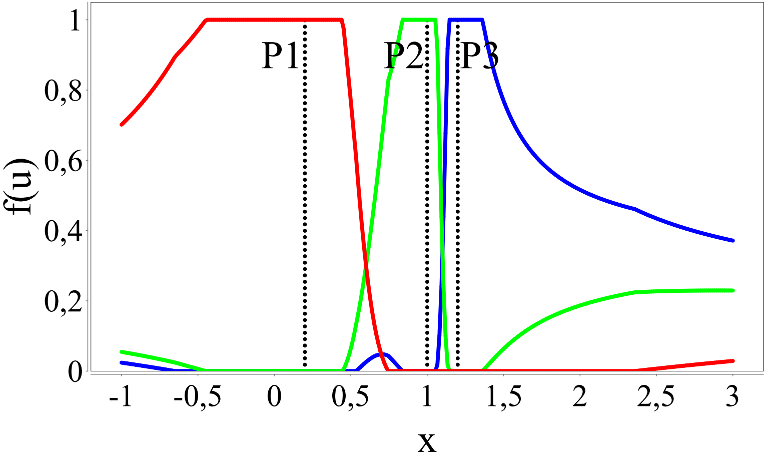
- $\boldsymbol{J}_{\text{PFCM}} = \sum_{i=1}^{c} \sum_{j=1}^{n} \left(\frac{1-\beta}{1+\beta} u_{ij}^2 + \frac{2\beta}{1+\beta} u_{ij} \right) d_{ij}^2$ with $d_{ij} = \|y_i - x_j\|$,

  $1 = \sum_{i=1}^{c} u_{ij}$, $u_{ij} > 1$ and $0 \geq \beta \geq 1$.
- with φ is a permutation that sorts the prototypes ascending by their distance and $\hat{c}$ is the number of clusters that have a positive membership value:

$$u_{ij} = \begin{cases} \frac{1}{1-\beta} \left(\frac{1+(\hat{c}-1)\beta}{\sum_{k=1}^{\hat{c}} \frac{d_{ij}^2}{d_{\varphi(k)j}^2}} - \beta \right) & \text{iff} \quad \varphi(i) \leq \hat{c} \\ 0 & \text{otherwise} \end{cases}$$

- in a 2 cluster environment: $u_{1j} = 1 \Leftrightarrow \frac{d_{1j}}{d_{2j}} < \beta$

PFCM fuzzified membership values example

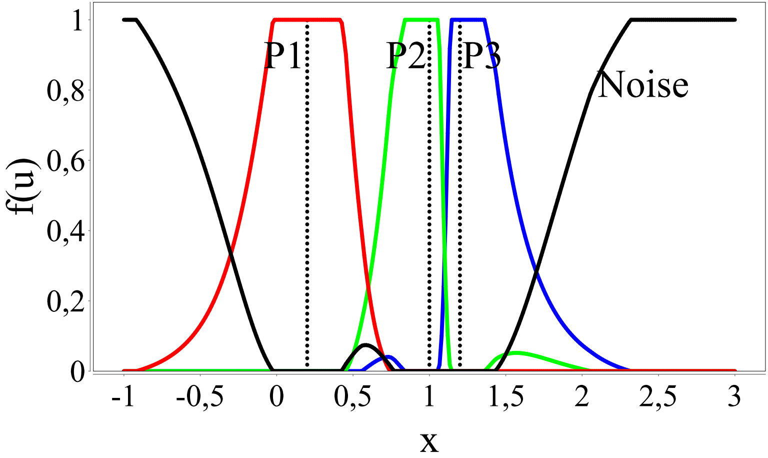


Noise fuzzy c-means with polynomial fuzzifier (PNFCM) [3]

- $\boldsymbol{J}_{\text{PNFCM}} = \sum_{i=0}^{c} \sum_{j=1}^{n} \left(\frac{1-\beta}{1+\beta} u_{ij}^2 + \frac{2\beta}{1+\beta} u_{ij} \right) d_{ij}^2$ with $d_{0j} = d_{\text{noise}} \geq 0$, $d_{ij} = \|y_i - x_j\|$ for $i = 1 \ldots c$, $1 = \sum_{i=1}^{c} u_{ij}$, $u_{ij} > 1$ and $0 \geq \beta \geq 1$.
- with φ and $\hat{c}$ as for PFCM:

$$u_{ij} = \begin{cases} \frac{1}{1-\beta} \left(\frac{1+(\hat{c}-1)\beta}{\sum_{k=0}^{\hat{c}} \frac{d_{ij}^2}{d_{\varphi(k)j}^2}} - \beta \right) & \text{iff} \quad \varphi(i) \leq \hat{c} \\ 0 & \text{otherwise} \end{cases}$$

PNFCM fuzzified membership values example



Maximal separated data objects on a hypersphere surface

- Sample many D-dimensional normal distributed data objects and project them to length 1
- The resulting data objects follow a rotation symmetric probability distribution
- Run Hard k-Means on the random generated data set
- Project prototypes on the hypersphere surface as new data objects

Data set properties

Dim	Clusters	Data Objects	per Cluster (Min - Max)	Angle [°] (Min)	Distance (Min - Max)
2	100	100000	811 - 3256	1.46	0.0253 - 2.0
5	100	100000	1800 - 2235	38.8	0.665 - 1.99
10	100	100000	1795 - 2237	54.1	0.910 - 1.95
20	100	100000	1831 - 2151	60.9	1.01 - 1.84
50	100	100000	1826 - 2162	74.0	1.20 - 1.74
100	100	100000	1853 - 2137	77.3	1.25 - 1.63
50	50	100000	3772 - 4192	80.3	1.29 - 1.60
50	100	100000	1839 - 2120	72.5	1.18 - 1.72
50	200	100000	901 - 1114	70.0	1.15 - 1.75
50	500	100000	309 - 480	63.0	1.04 - 1.77